

Machine Learning 1

Lecture 2.3 - Maximum Likelihood

Erik Bekkers

(Bishop 1.2.3 - 1.2.5)



Maximum Likelihood Principle

- ▶ Dataset $D = (x_1, x_2, \dots, x_N)$ of N independent observations.
- ▶ Likelihood of the dataset: $p(D|\mathbf{w})$
- ▶ Maximum likelihood principle: the most likely “explanation” of D is given by \mathbf{w}_{ML} which maximizes the likelihood function

$$\mathbf{w}_{\text{ML}} = \underset{\mathbf{w}}{\text{argmax}} p(D|\mathbf{w})$$

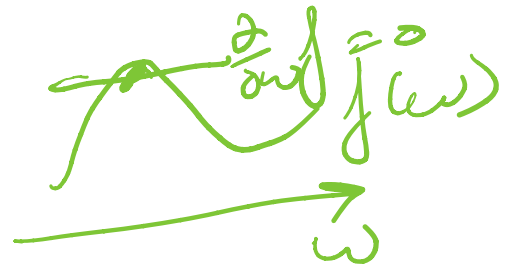
- ▶ i.i.d. assumption: each $x_i \in D$ is independently distributed ^{identically} according to the same distribution, conditioned on \mathbf{w} .

$$x \sim p(x|\mathbf{w})$$

- ▶ If i.i.d., joint distribution

$$p(D|\mathbf{w}) = p(x_1, x_2, \dots, x_N|\mathbf{w}) = \prod_{i=1}^N p(x_i|\mathbf{w})$$

Maximum Likelihood Estimation



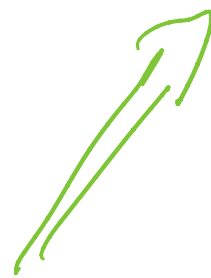
- Maximum likelihood estimation \mathbf{w}_{ML}

$$\mathbf{w}_{\text{ML}} = \arg \max_{\mathbf{w}} p(D|\mathbf{w}) = \arg \max_{\mathbf{w}} \prod_{i=1}^N p(x_i|\mathbf{w})$$

$$\Rightarrow \boxed{\frac{\partial}{\partial \mathbf{w}} p(D|\mathbf{w}) = 0}$$

numerical underflow/overflow

- How do we maximize?



$$\frac{\partial}{\partial \mathbf{w}} \log p(D|\mathbf{w}) = 0$$

$$\frac{1}{p(D|\mathbf{w})} \left\{ \frac{\partial}{\partial \mathbf{w}} p(D|\mathbf{w}) = 0 \right.$$

- Maximize log-likelihood instead:

$$\mathbf{w}_{\text{ML}} = \arg \max_{\mathbf{w}} \prod_{i=1}^N p(x_i|\mathbf{w}) = \arg \max_{\mathbf{w}} \log \prod_{i=1}^N p(x_i|\mathbf{w})$$

$$= \arg \max_{\mathbf{w}} \sum_{i=1}^N \log p(x_i|\mathbf{w})$$

- Error function: $E(D; \mathbf{w}) = -\log p(D|\mathbf{w}) = -\sum_{i=1}^N \log p(x_i|\mathbf{w})$

ML Estimator for Gaussian Distributions (I)

- ▶ i.i.d. Gaussian distributed real variables $D = (x_1, x_2, \dots, x_N)$

$$p(x|\mathbf{w}) = \mathcal{N}(x|\mu, \sigma^2) \quad \longrightarrow \quad p(D|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^N \exp \left[-\frac{1}{2\sigma^2} (x_i - \mu)^2 \right]$$

$$p(x_1, x_2, \dots, x_N | \mu, \sigma^2) = p(x_1 | \mu, \sigma^2) p(x_2 | \mu, \sigma^2) \dots$$

- ▶ Log likelihood

$$\begin{aligned} \log p(D|\mu, \sigma^2) &= \log (2\pi\sigma^2)^{-N/2} + \sum_{i=1}^N \log \exp \left[-\frac{1}{2\sigma^2} (x_i - \mu)^2 \right] \\ &= -\frac{N}{2} \log (2\pi\sigma^2) + \sum_{i=1}^N -\frac{1}{2\sigma^2} (x_i - \mu)^2 \end{aligned}$$

- ▶ Estimate model parameters: $\mu_{ML}, \sigma_{ML}^2 = \underset{\mu, \sigma^2}{\operatorname{argmax}} \log p(D|\mu, \sigma^2)$

$$\frac{\partial}{\partial \mu} \log p(D|\mu, \sigma^2) = 0$$

solve for $\mu \rightarrow \mu_{ML}$

ML Estimator for Gaussian Distributions (II)

- log likelihood:

$$\log p(D|\mu, \sigma^2) = -\frac{N}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^N (x_i - \mu)^2$$

$$\frac{\partial}{\partial \mu} (x_i - \mu)^2 = 2(x_i - \mu)$$

- Maximum Likelihood solution for μ

$$\frac{\partial}{\partial \mu} \log p(D|\mu, \sigma^2) = \frac{1}{\cancel{2}\sigma^2} \sum_{i=1}^N \cancel{2} (x_i - \mu) = 0$$

$$\Rightarrow \sum_{i=1}^N (x_i - \mu) = 0$$

$$\Rightarrow \sum_{i=1}^N \mu = \sum_{i=1}^N x_i$$

$$\Rightarrow N \cdot \mu = \sum_{i=1}^N x_i \Rightarrow$$

sample mean

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

ML Estimator for Gaussian Distributions (II)

- log likelihood:

$$\log p(D|\mu, \sigma^2) = -\frac{N}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^N (x_i - \mu)^2$$

- Maximum Likelihood solution for σ^2

$$\frac{\partial}{\partial \sigma^2} \log p(D|\mu, \sigma^2) = -\frac{N}{2} \frac{1}{\cancel{2\pi}\sigma^2} \cdot \cancel{2\pi} + \frac{1}{2\sigma^4} \sum_{i=1}^N (x_i - \mu)^2 = 0$$

$\frac{\partial}{\partial \sigma^2} \cdot \frac{1}{2\sigma^2} = -\frac{1}{2\sigma^4}$

$$\Rightarrow -N\sigma^2 + \sum_{i=1}^N (x_i - \mu)^2 = 0$$

$\cdot 2\sigma^4$

$$\Rightarrow \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad \text{sample variance}$$

$-\frac{1}{N}$

ML Estimator for Gaussian Distributions (IV)

- How well do the ML estimators represent the true parameters?

$$p(D|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^N \exp \left[-\frac{1}{2\sigma^2} (x_i - \mu)^2 \right]$$

- If I draw multiple datasets, what is the expected value of μ_{ML} ?

$$D_1 = \{x_1, \dots, x_n\}, \quad D_2 = \{x_1, \dots, x_n\}$$

- ML estimate of the mean:


$$\begin{aligned} \mathbb{E}_{D \sim p(D|\mu, \sigma^2)} [\mu_{\text{ML}}] &= \mathbb{E} \left[\frac{1}{N} \sum_{i=1}^N x_i \right] = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{D \sim p(D|\mu, \sigma^2)} [x_i] \\ &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{x_i \sim p(x_i|\mu, \sigma^2)} [x_i] = \frac{1}{N} \sum_{i=1}^N \mu = \mu \end{aligned}$$

- Bias of estimator:

$$\mathbb{E}[\mu_{\text{ML}}] - \mu = 0$$

ML Estimator for Gaussian Distributions (V)

- ▶ ML estimate of the variance:

$$\mathbb{E}_{D \sim p(D|\mu, \sigma^2)}[\sigma_{ML}^2] = \mathbb{E} \left[\frac{1}{N} \sum_{i=1}^N \left(x_i - \frac{1}{N} \sum_{n=1}^N x_n \right)^2 \right]$$


ML Estimator for Gaussian Distributions (V)

- ▶ ML estimate of the variance:

$$\mathbb{E}_{D \sim p(D|\mu, \sigma^2)}[\sigma_{ML}^2] = \mathbb{E} \left[\frac{1}{N} \sum_{i=1}^N \left(x_i - \frac{1}{N} \sum_{n=1}^N x_n \right)^2 \right]$$

$$= \frac{1}{N} \sum_{i=1}^N \mathbb{E} \left[\left(x_i - \frac{1}{N} \sum_{n=1}^N x_n \right)^2 \right]$$

ML Estimator for Gaussian Distributions (V)

- ▶ ML estimate of the variance:

$$\mathbb{E}_{D \sim p(D|\mu, \sigma^2)}[\sigma_{ML}^2] = \mathbb{E} \left[\frac{1}{N} \sum_{i=1}^N \left(x_i - \frac{1}{N} \sum_{n=1}^N x_n \right)^2 \right]$$

$$= \frac{1}{N} \sum_{i=1}^N \mathbb{E} \left[\left(x_i - \frac{1}{N} \sum_{n=1}^N x_n \right)^2 \right] = \frac{1}{N} \sum_{i=1}^N \mathbb{E} \left[x_i^2 - \frac{2x_i}{N} \sum_{n=1}^N x_n + \frac{1}{N^2} \sum_{m=1}^N \sum_{n=1}^N x_m x_n \right]$$

ML Estimator for Gaussian Distributions (V)

- ML estimate of the variance:

$$\begin{aligned} \mathbb{E}_{D \sim p(D|\mu, \sigma^2)}[\sigma_{ML}^2] &= \mathbb{E} \left[\frac{1}{N} \sum_{i=1}^N \left(x_i - \frac{1}{N} \sum_{n=1}^N x_n \right)^2 \right] \\ &= \frac{1}{N} \sum_{i=1}^N \mathbb{E} \left[\left(x_i - \frac{1}{N} \sum_{n=1}^N x_n \right)^2 \right] = \frac{1}{N} \sum_{i=1}^N \mathbb{E} \left[x_i^2 - \frac{2x_i}{N} \sum_{n=1}^N x_n + \frac{1}{N^2} \sum_{m=1}^N \sum_{n=1}^N x_m x_n \right] \\ &= \frac{1}{N} \sum_{i=1}^N \left\{ \mathbb{E}[x_i^2] - \frac{2}{N} \sum_{n=1}^N \mathbb{E}[x_i x_n] + \frac{1}{N^2} \sum_{m=1}^N \sum_{n=1}^N \mathbb{E}[x_m x_n] \right\} \end{aligned}$$

$$\begin{aligned} &= \frac{1}{N} \sum_{i=1}^N \left\{ \mu^2 + \sigma^2 - \frac{2}{N} (N\mu^2 + \sigma^2) + \frac{1}{N^2} (N^2\mu^2 + N\sigma^2) \right\} \\ &= \cancel{\mu^2} + \sigma^2 - 2\cancel{\mu^2} - \frac{2}{N}\sigma^2 + \cancel{\mu^2} + \frac{1}{N}\sigma^2 = \sigma^2 - \frac{1}{N}\sigma^2 = \frac{N-1}{N}\sigma^2 \end{aligned}$$

$$\mathbb{E}[x_i x_j] = \begin{cases} \mu^2 + \sigma^2 & \text{if } i = j \\ \mu^2 & \text{if } i \neq j \end{cases}$$

$\text{Cov}[x_i, x_i] = \mathbb{E}[x_i^2] - \mathbb{E}[x_i]^2 = \sigma^2$
 $\text{Cov}[x_i, x_j] = \mathbb{E}[x_i x_j] - \mathbb{E}[x_i] \mathbb{E}[x_j] = 0$

ML Estimator for Gaussian Distributions (VI)

- ▶ For data generated from

$$p(D|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^N \exp \left[-\frac{1}{2\sigma^2} (x_i - \mu)^2 \right]$$

- ▶ ML gives biased estimator

$$\mathbb{E}[\sigma_{\text{ML}}^2] = \frac{N-1}{N} \sigma^2$$

- ▶ Unbiased variance estimator:

$$\tilde{\sigma}^2 = \frac{N}{N-1} \sigma_{\text{ML}}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2$$

$$\mathbb{E}[\tilde{\sigma}^2] = \sigma^2$$

Biased Maximum Likelihood Estimator

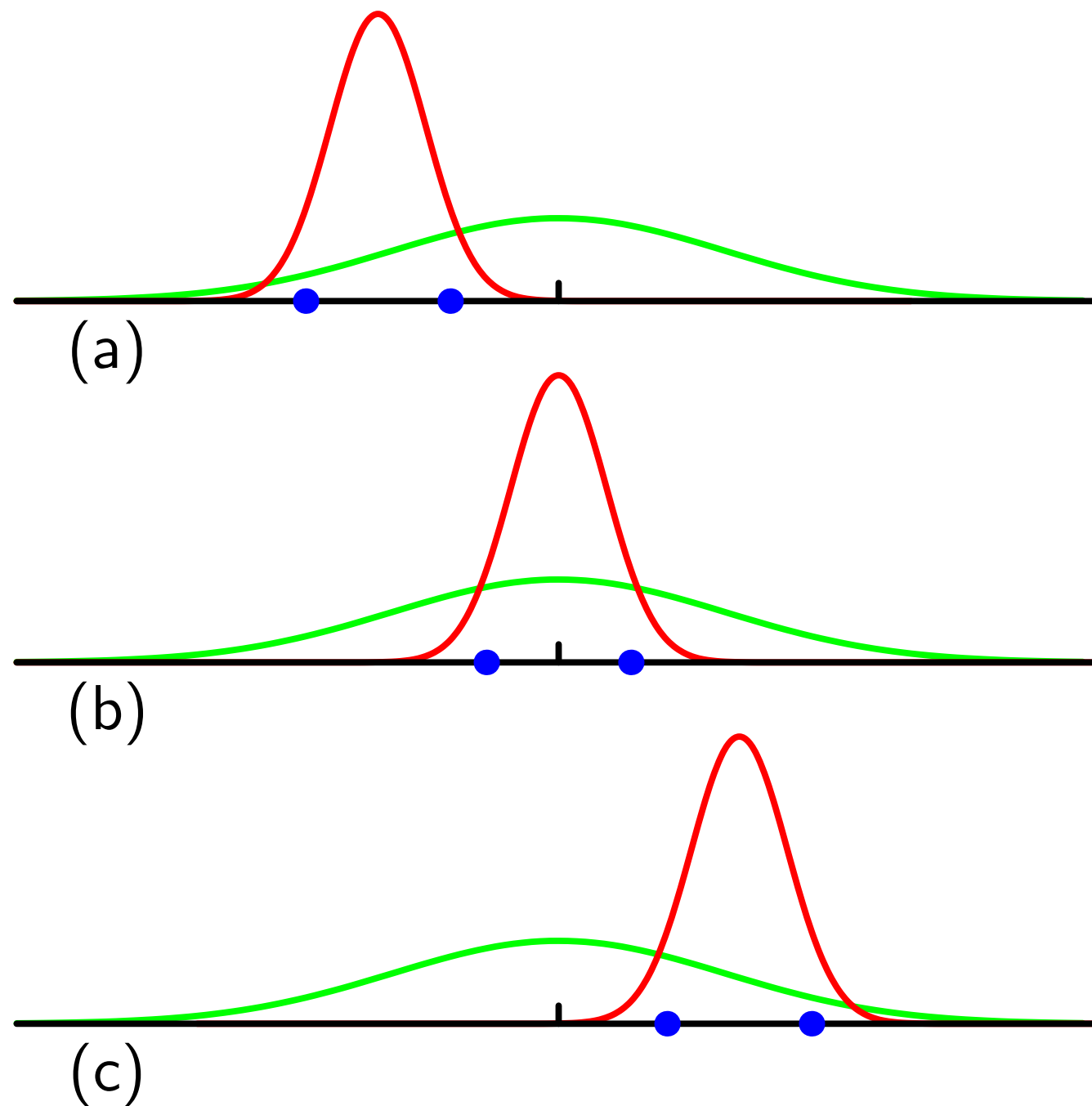


Figure: Bias in ML estimator for variance (Bishop 1.15)